情 報 学 基 礎　　68－13
自 然 言 語 処 理 151－13
（２００２．９．１８）

# 文献クラスタの概念的特徴づけを用いた文献の自動分類

中島　誠　　伊藤　哲郎
大分大学　知能情報システム工学科
{nakasima, ito}@csis.oita-u.ac.jp

大量の電子化された文献の中から必要とするものをうまく取り出すための方策として，文献を内容に応じてクラスタに分類しておくことが重要とされてきた．従来からの自動分類の手法の多くは，文献キーワードの出現頻度をもとに，未分類の文献を既存のクラスタに精度高く分類できるよう方法を求めてきた．しかしながら，望む文献をうまく取り出すには，文献の管理に際し各文献クラスタの内容や他のクラスタとの関連を利用者が容易に理解できる表現が望まれる．この要求に応えるため，ここでは，キーワードをシソーラス等の概念階層中の記述子に置き換え，分類精度の向上に寄与しながら，クラスタの内容を概念的に特徴づける簡潔な表現が得られる自動分類の手法を定式化する．実験を通じて，従来の方法と遜色のない分類精度をより簡潔な表現で得られることを確かめた．

# Automated Document Classification based on Conceptual Characterizing of Document Clusters

Makoto NAKASHIMA and Tetsuro ITO
Department of Computer Science and Intelligent Systems
Oita University

The categorization of documents into predefined clusters becomes increasingly important due to the increased availability of documents in digital form. The keyword-based approaches in automated categorization of documents are insufficient in clarifying the contents of the clusters, since the keywords usually have some conceptual relations. We here formulate a document classification method of finding *simplified conceptual expressions* based on the subject descriptors in a concept hierarchy for characterizing the clusters so as to clarifying the contents of documents in each cluster by keeping the classification accuracy fairly high. The simplification is done by removing the less informative descriptors and by evaluating the changed expressions based on the classification accuracy when any document in the predefined clusters is treated as a new document. The availability of the proposed method was also examined computationally.

## 1. はじめに

近年，計算機上での文献の作成が当たり前になるとともに，WWW のようなネットワーク上での情報提供技術の普及により，大量の電子化された文献が利用可能となった．これらを管理する有効な手段の一つとして，文献を扱う内容に応じて分類しておくことがこれまで以上に重要となっている．実際，Yahoo [15]や Excite [6]などの WWW 上の検索エンジンでは, Web ページがその内容に従ってクラスタ化され管理されている．また，計算機上のファイルや電子メールの管理にも応用されている．

文献の自動分類の研究は，人手により文献を既存のクラスタに分類する負担を軽減することを目的の一つとして行われてきた．これらの多くは，文献中でのキーワードの出現頻度をもとに，文献がどのクラスタに属する確率が高いかを求めることにより分類を行う．精度の高い分類結果を得ることが重要とされてきた．しかしながら，キーワードは互いに概念的な関連をもち, 頻度を基にキーワードの重要性を表すのは難しい．関連性を捉えるためにキーワード同士の共起頻度を考慮に入れる考えもあるが，キーワードのリストから，それらが何を意味するのかを理解し文献クラスタの内容を把握するのは難しく，他のクラスタとの関連も明

確でない.

分類された文献のうち必要なものをうまく取り出すには，文献クラスタの内容が理解し易い形で表現されることが必要である [4]．ここでは，この要求に応えるために，既知のクラスタに属する文献が与えられたとして，クラスタと文献の概念的表現を基にした自動分類の方法について議論する．文献の表現には，様々な分野で各文献の内容を捉えるために整備された概念階層としてのシソーラスあるいは分類体系にある記述子(概念を示すための語句あるいは記号)を利用する．階層中では, 記述子間の概念的関連が規定されている．このような状況下で,

(i) 文献をできるだけ精度高く分類できながら
(ii) 概念的に簡潔に内容を捉えられる

ように文献クラスタを特徴づける表現を得る方法を定式化する.

分類の基本的な流れは，与えられた文献クラスタを学習用データとして，クラスタの表現を求め，これをもとに, 未分類の文献の属すべきクラスタを決定する．ここでは，各文献のキーワードを概念階層中の記述子に置き換え，文献の概念的な内容表現とみなす．各クラスタの表現は，クラスタに属する文献の表現の記述子を集めたものとなる.

クラスタの表現の簡略化は，クラスタに属する文献に共通する記述子ほどクラスタの内容を捉えるために必要性が高いとし，必要性の低いものから順に除去しながら，分類精度の向上に役立つと考えられる記述子を見つけることで行う. ここで, 表現の良さの評価は，各クラスタに属する文献それぞれを分類対象の文献とし，他の文献との概念的関連性によって属すべきクラスタを決定したときの分類精度により行う．文献間の概念的関連性は，それぞれの表現にある各記述子で概念階層中で近くにあるものを対応づけて捉える [9]．より近くの記述子がより多く対応させられているほど関連性が強いとし，分類対象の属すべきクラスタは，関連性の強い文献が多く含まれているクラスタとする.

実際に文献を分類して管理する場合，最終的な文献の分類先は人手によって決定される事が多い．例えば Yahoo では, Web ページをどのクラスタに含めるかは，ユーザが必要なページを効果的に発見できるよう専門のスタッフによって決定される．計算機上のファイルの整理もまた，ユーザ自身の判断に従い行われることが多い．これらの作業をうまく行うには，既存のクラスタや中の文献の内容が，分類対象の文献との関連が捉えやすいように表されている必要がある．提案した方法によれば，概念階層上での記述子間の関連をみながらクラスタと分類対象の文献との関連を理解でき，高い精度で分類対象が属すべきクラスタの示唆も可能となる.

提案した方法の有効性は, ACM の Transactions に掲載された文献と OHSUMED データベースに含まれる医学論文誌に掲載された文献を集めた 2 つのコレクションを対象に調べた．それぞれのコレクションでは，ACM Computing Classification System (CCS) [1] と Medical Subject Heading (MeSH) [10] が概念階層として利用できる．各論文誌には，共通の話題を扱う論文が掲載され，各論文誌の文献がクラスタを成すとして扱った. 何れのコレクションでも, 簡潔でありながら，分類精度の高いクラスタの表現が得られることが分かった.

## 2. 文献の自動分類

自動分類の過程は，既存の文献クラスタの特徴づけを行う学習段階と，特徴づけた結果をもとに，未分類の文献が属すべきクラスタを決める分類段階からなる. 従来の方法では，機械学習の技法を利用したものが多く，中でも，k-最近隣(k-NN)法 [16],[17]および Support Vector Machine(SVM) [7],[5]が優れた評価を得ている[17]．前者は，学習段階で既存のクラスタに属する文献を蓄積し，未分類の文献の属すべきクラスタは，クラスタごとに蓄積された文献との類似度の大きさに従って決定される．後者は，各クラスタに属する文献と他の文献を分離する，超空間上での各文献との距離が最大の分離超平面を求める学習方法である．分類段階では，基本的に各クラスタの文献で分離超平面に近いものとの類似度に従って，未分類の文献が属すべきクラスタを決定する.

上の方法に代表されるように，従来からの方法の多くは，キーワードによる文献の表現を用い，未分類の文献を如何に精度高く分類するかを問題にしていた．それ故，クラスタの特徴づけは，異なるクラスタの文献を区別するためのもので，文献の属すべきクラスタが決定されたとして，どのような内容の文献を含むクラスタに決められたかを理解することは難しい. また，他のクラスタとの関連も捉えることが困難である.

以下では，既存の文献クラスタとそれに属する文献が与えられたとして, 高い分類精度を導くだけでなく，その内容を捉えやすい，簡潔な概念的表現で文献クラスタを特徴づける自動分類の方法について述べる．各

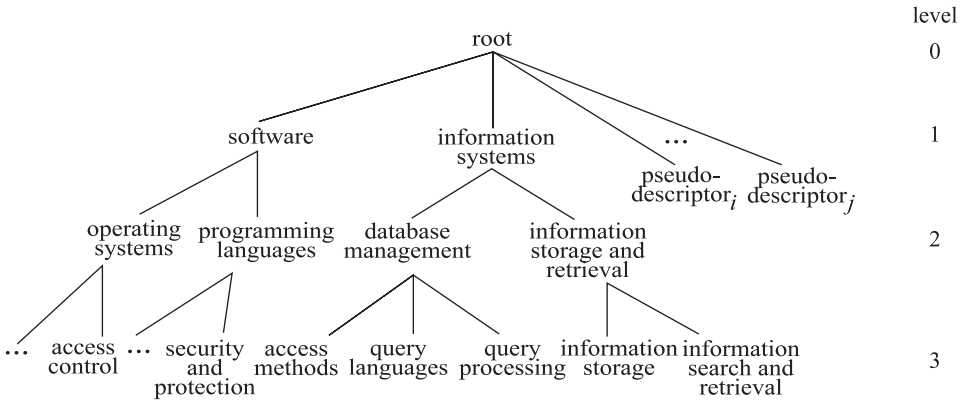


図 1. 概念階層（ACM の CCS の一部）

クラスタやこれに属する文献の表現には，文献のキーワードを概念階層中の記述子と置き換えて用いる．そして，各クラスタに共通する記述子を残しながら不必要な記述子を除去して，クラスタを区別するだけでなく，内容を特徴づけるクラスタの表現を得る．未分類の文献の属すべきクラスタは，概念階層上での記述子間の関連をもとに，既存のクラスタに属する文献との概念的な関連性の強さに従って決定される．これにより，内容的に似た文献を効果的に分類できるようになる．また，クラスタの表現と分類対象の文献の表現を参照することで，どのような関連でクラスタが決定されたかも明確になる．

## 3. 概念的表現

今，4 つの文献 $d_1$ から $d_4$ が 2 つのクラスタ $C_1$ と $C_2$ に分けられているとする．各文献には，以下のように重みつきキーワードがつけられているとする．

$d_1$ : [ (storage: 0.30), (query: 0.30), (database: 0.20),
(access: 0.20)],
$d_2$ : [ (retrieval: 0.40), (search: 0.40), (language: 0.20) ],
$d_3$ : [ (security: 0.50), (access: 0.30), (language: 0.20) ],
$d_4$ : [ (security: 0.40), (programming: 0.30),
(management: 0.30) ].

このようなキーワードによる表現では，$C_1$ や $C_2$ それぞれの全体の内容やこれらの関連性を簡単に捉えることが難しい．ここでは，文献の概念的な扱いのためにキーワードを部分文字列として含む, 概念階層中の記述子に置き換える．置き換えられた記述子のリストを文献の概念的表現と呼ぶ.

図 1 に概念階層の例を示す．階層中では，低い（高い）レベルにおかれた記述子ほど上位（下位）の概念を示す．例えば，記述子“ information storage and retrieval ”は“ information storage ”より一般的であるといい，後者は前者より特定的であるともいう．また,“ information systems ”も“ information storage ”より一般的である．すべての記述子より一般的なものを“ root ”とし，レベル 0 におく．ここで例えば,“ information storage ”に対し“ information search and retrieval ”との最小共通祖先“ information storage and retrieval ”は，他の記述子とのそれよりも特定的であり，“ information search and retrieval ”は“ information storage ”に概念的により近くにあるという．

図 1 より文献 $d_1$，$d_2$，$d_3$，$d_4$ の概念的表現 $d_1^c$，$d_2^c$，$d_3^c$，$d_4^c$ は次のようになる．

$d_1^c$ : [ (information storage and retrieval: 0.15),
(information storage: 0.15), (query languages: 0.15),
(query processing: 0.15), (database management: 0.20),
(access control: 0.10), (access methods: 0.10) ],
$d_2^c$ : [ (information storage and retrieval: 0.20),
(information search and retrieval: 0.60),
(programming languages: 0.10), (query languages: 0.10)],
$d_3^c$ : [ (security and protection: 0.50), (access control: 0.15),
(access methods: 0.15), (programming languages: 0.10),
(query languages: 0.10) ],
$d_4^c$ : [ (security and protection: 0.40),
(programming languages: 0.30),
(database management: 0.30) ].

各記述子の重みは，キーワードの重みを置き換えた記述子の数で割った値を，同じ記述子ごとに合計して求めた．キーワードを部分文字列として含む記述子がない場合は，そのようなキーワードごとに作られた記述子“ pseudo descriptor ”に置き換えるようにする.“ pseudo descriptor ”は

図 1 に示すように自身を除いてより特定的な記述子を持たない.

今，ある記述子が文献の内容を表すのに必要とされるならば，概念階層中でそれと近くにある記述子も同じく必要であると言える．このことを考慮して，記述子を概念的にまとめた形の表現を扱う．具体的には，各表現を，概念階層中で上位のレベルにある記述子を用いた表現を用いる．上の例で，レベル 2 にある記述子で書き直した文献の表現は次となる．

$g(d_1^c,2)$ : [(information storage and retrieval: 0.30),
(database management: 0.60),
(operating systems: 0.10)]
$g(d_2^c,2)$ : [ (information storage and retrieval: 0.80),
(programming languages: 0.10),
(database management: 0.10)],
$g(d_3^c,2)$ : [ (programming languages: 0.60),
(operating systems: 0.15)
(database management: 0.25)]
$g(d_4^c,2)$ : [(programming languages: 0.70),
(database management: 0.30)]

ここで，例えば，$g(d_1^c,2)$ にある記述子 "database management"の重み 0.60 は, この記述子より特定的な $d_1^c$ の記述子の重みを合計して求めた.

上のように書き直した各文献の概念的表現の記述子をクラスタごとに集めて，クラスタの概念的表現とする．各記述子の重みは，クラスタ中の文献の概念的表現への出現確率とする. クラスタ $C_1$ と $C_2$ の概念的表現 $G(C_1,2)$ と $G(C_2,2)$ は次となる.

$G(C_1,2)$ : [(information storage and retrieval: 1.00)
(database management: 1.00),
(operating systems: 0.50),
(programming languages: 0.50)]
$G(C_2,2)$ : [ (programming languages: 1.00),
(database management: 1.00),
(operating systems: 0.50),]

## 4. 簡潔なクラスタの表現

クラスタの概念的表現は，中に含まれている文献の内容を特徴づけるのに十分な記述子を含んでいると言える．しかしながら，全ての記述子が必要であるとは限らない．クラスタの表現の中で重みが大きい記述子は多くの文献に現れることからクラスタを特徴づけるのに必要性が高いと言える．これらは異なるクラスタを区別するのに必要とは限らなくとも，クラスタに属する文献の内容を理解したり，クラスタ間の関連を捉えるのに必要である. 以上を考慮し，重みの低い記述子から順に除去していき，文献を高い精度で分類しながら, かつ簡潔なクラスタの表現を求めていく.

どれだけの記述子を除去したら良いかを決めるには，得られた表現が 1.で述べた(i)と(ii)を満たすかどうかの評価が必要になる．評価が向上したりあるいは変化がなければ，新たな表現は元の表現の簡潔な表現と言える．**3.**の例を用いて，クラスタの簡潔な表現を求める方法を説明しよう.

まず，各文献 $d_1$，$d_2$，$d_3$，$d_4$ を分類対象として捉える．そして, 最初に記述子を選ぶために, 重みの閾値を設定し，これ以下の重みをもつ記述子を $G(C_1,2)$ と $G(C_2,2)$ より取り除いた表現をそれぞれ $G'(C_1,2)$ と $G'(C_2,2)$ とする．これらより特定的な記述子のみで文献を表現する．次に，各分類対象への概念的な近さをもとに他の文献をランキングし，全ての分類対象にわたっての分類精度を計算する．分類精度が低下しなければ，$G'(C_1,2)$ と $G'(C_2,2)$ を $G(C_1,2)$ と $G(C_2,2)$ それぞれの簡潔な表現と捉えることができる．このプロセスを閾値を増加させながら繰り返す.

2 つの文献間の概念的関連性を捉えるには，これらの概念的な表現をもとに，両者について述べた一般的な表現を利用する．例えば $d_1^c$ と $d_2^c$ の一般的な表現 $d_{12}^c$ としては，$d_1^c$ と $d_2^c$ の各記述子について，それより一般的な記述子が $d_{12}^c$ に含まれ，また，$d_{12}^c$ の各記述子について，それより特定的な記述子が $d_1^c$ および $d_2^c$ に含まれる表現とする[9].

このような表現は, $d_1^c$ ($d_2^c$)の各記述子を $d_2^c$ ($d_1^c$)の任意の記述子に対応させてそれらの最小共通祖先を求める．そして，それら最小共通祖先のうちの最も特定的なものを $d_{12}^c$ の 1 つの記述子とする．例えば，$d_1^c$ の " information storage " は，$d_2^c$ の " information search and retrieval " (あるいは " information storage and retrieval " )と対応させられ，$d_{12}^c$ の " information storage and retrieval"が求まる．記述子の重みは，対応づけられた 2 つの記述子の重みの小さい方と定める．同じ記述子ごとに重みを合計すると，$d_{12}^c$ は次のようになる.

$d_{12}^c$ = [ (information storage and retrieval: 0.60),
(query languages: 0.20), (database management: 0.30),
(software: 0.20) ]

$d_1$ と $d_2$ の関連性の強さは $d_{12}^c$ に含まれる記述子の特定さを合計して測る．以後この値を $d_1$ と $d_2$ の類似度と呼び $S_c(d_1,d_2)$ と書く. 特定さとしてここでは, 文献 [9]を参考に，レベル $l$ の記述子に，$l/h$ を割り当てるようにする（$h$ は最も高いレベルを表す）．例では，" query languages " に 1,

“information storage and retrieval”と“database management”に2/3,“software”に1/3が割り当てられる．$S_c(d_1,d_2)$は0.43 (=(0.60·2/3+0.20·1+0.30·2/3+0.20·1/3)/2)となる．ここで，分母の2は正規化のための定数である．

$G(C_1,2)$と$G(C_2,2)$を用いて表現の簡略化の例を示そう．最初に閾値は0に設定しておく．文献$d_2$，$d_3$，$d_4$を$d_1$との類似度に従ってランキングした結果を表1の上半分に示す．

表1. 概念的関連性に従うランキング結果

| 閾値 | 分類対象 | | | |
|---|---|---|---|---|
| | $d_1$ | $d_2$ | $d_3$ | $d_4$ |
| 0.00 | $d_3$ (0.48) | $d_1$ (0.43) | $d_4$ (0.69) | $d_3$ (0.69) |
| | $d_2$ (0.43) | $d_3$ (0.30) | $d_1$ (0.48) | $d_1$ (0.36) |
| | $d_4$ (0.36) | $d_4$ (0.25) | $d_2$ (0.30) | $d_2$ (0.25) |
| 0.50 | $d_2$ (0.44) | $d_1$ (0.44) | $d_4$ (0.69) | $d_3$ (0.69) |
| | $d_3$ (0.39) | $d_3$ (0.20) | $d_1$ (0.39) | $d_1$ (0.35) |
| | $d_4$ (0.35) | $d_4$ (0.16) | $d_2$ (0.20) | $d_2$ (0.16) |

既存のクラスタに属する文献との類似度から分類対象の属すべきクラスタを決定するのに，一般にk-NN法が有効とされる [17]．ここでは，分類対象に対する，ランキング上位k個の文献の類似度を，これらが属するクラスタごとに合計し，最大の値を得たクラスタと決める．上記の例では，各分類対象以外で同じクラスタに属する文献の数が1であることから，k=1とする．例えば，$C_1$にある$d_1$を分類対象とした時，同じクラスタにある$d_2$は2位にランクされてしまい，分類対象全体の分類精度は0.75となる．

閾値を0.5とすると，$G'(C_1,2)$と$G'(C_2,2)$は次となる．

$G'(C_1,2)$: [(information storage and retrieval:1.00),
(database management: 1.00) ],
$G'(C_2,2)$ : [(programming languages: 1.00),
(database management: 1.00) ]

例えば，$G'(C_1,2)$を参照することで，$d_1^c$の中で“information storage and retrieval”より特定的な“information storage and retrieval”と“information storage”，および“database management”より特定的な“query languages”，“query processing”，“database management”と“access methods”が$d_1$の概念的表現に用いられる（記述子の重みは正規化する）．

閾値0.5のときのランキング結果を表1の下半分に示す．各分類対象に対して，同じクラスタの文献はすべて1位にランクされている．$G'(C_1,2)$と$G'(C_2,2)$を参照することで，分類精度が向上していると言える．また，クラスタ間の関連も$G(C_1,2)$と$G(C_2,2)$より明確になっている．$C_1$と$C_2$に属する全ての文献がそれぞれ，“information storage and retrieval”と“programming languages”に関するものでありながら，双方のクラスタとも，”database management”に関する文献を含んでいる．

クラスタの概念的表現を簡略化する方法を手続きの形でまとめる．

[手続き1]
:文献クラスタの集合$\{C_1,\ldots,C_j,\ldots\}$と各クラスタ$C_j$ ($j$≧1)に属する文献が与えられているとする．各$C_j$についてレベル$l$の記述子を用いて$G(C_j,l)$を求める．閾値の初期値を0とする．
(S1) 分類精度が向上するか一定の間，閾値を増加させながらS1.1とS1.2を行う．
(S1.1) 各$G(C_j,l)$から重みが閾値以下の記述子を除去して$G'(C_j,l)$を得る．
(S1.2) 各$C_j$の文献の表現で$G'(C_j,l)$にある記述子より特定的な記述子を用いた場合の分類精度を求める．
(S2) S1の結果得られた各$G'(C_j,l)$を簡潔なクラスタ表現として出力する．

ところで，文献の概念的表現にある記述子は，キーワードを置き換えて得たもので，すべてが文献の内容を表すのに必要とは限らない．ここでは，クラスタの表現の簡略化の前処理として，各文献の表現から，文献内容を表すのに必要性の低いものを除去しておくようにする．この処理は，上の手続き1で，クラスタ表現の代わりに各文献についての表現$g(d_1^c,l)$を対象にし，各文献を質問と捉えて，他の文献を概念的関連性に従ってランキングし，情報検索での評価測度（再現率，適合率）を文献の表現の評価に用いることで同様に行うことができる [9]．質問となった文献と同じクラスタに属する文献を適合文献と捉える．

## 5. 未分類文献の分類

未分類文献の属すべきクラスタは，4.で簡潔なクラスタ表現を求めるためにも行ったように，既存のクラスタに属する文献との類似度に従って決定される．このとき，クラスタに属する文献の表現は，手続き1で求めたクラスタの表現を参照して得る．文献の分類の手続きを以下にまとめておく．

[手続き2]
:各文献クラスタ$C_j$ ($j$≧1)についてレベル$l$の記述子を用いた簡潔なクラスタ表現$G'(C_j,l)$が求められているとする．未分類文献$d_x$が与えられたとする．

(S1) $d_x$の概念的表現$d_x^c$を求める．

(S2) 各$C_j$の文献を，$G'(C_j,l)$ にある記述子より特定的な記述子のみを用いた表現をもとにした$d_x$との類似度に従いランキングする．

(S3) ランキング上位k個の文献の類似度を文献の属するカテゴリごとに合計する．値が最大のカテゴリを$d_x$の分類先とする．

## 6. 実験

定式化した方法の有効性を調べるために ACM Digital Library [2]にある 5 つの論文誌に掲載された 400 編, および OHSUMED [11]にある 10 の論文誌に掲載された 500 編からなる 2 つの文献コレクションを用意した（それぞれ, ACM, OMD と略す）．各コレクションの論文誌名と文献数を表 2 と 3 に示す．同じ論文誌に掲載された文献の集まりそれぞれを 1 つのクラスタとして捉える．概念階層として，ACM では，CCS [1]を，OMD では MeSH [10]を利用した．各文献のキーワードは，それぞれの表題とアブストラクト中の高頻度語・句を自動抽出し, 語幹を取り出しておいた [12]. 各キーワードの重みは，$tf \times (1+\log(M/df))$により求めた．ここで，$tf$ はキーワードの出現頻度（文献ごとのキーワード頻度の合計で正規化した）．$M$ は文献の数である．各文献の概念的表現は，コレクションごとの概念階層中の記述子を利用して求めた．ACM(OMD)の各文献のキーワード数, 概念的表現中の記述子の数およびレベル 2 での表現$g(d_i^c,2)$の記述子の数の平均は，それぞれ 16.1(25.9)，52.3(149.0), 15.5(39.0)であった．

各コレクションに対し，それぞれの論文誌から 40%の文献を取り出し，学習用コレクションとして用い，残りをテスト用コレクションとして用いた．各学習用コレクションからはさらに，それぞれの論文誌ごとに全体の 10%, 20%, 30%の割合で文献を取り出し，異なるサイズの学習用コレクションでの有効性も調べた．前処理として学習用コレクションの文献の表現から$g(d_i^c,2)$を用いて，必要となる可能性の低いものを除去しておいた．評価基準として，検索効率の評価にもちいられる再現率と適合率をあわせた $F_1$ 測度 [14]を使った．効率が維持されるか向上する間，各表現中の記述子を重みの低い順に除去した．ACM(OMD)では，平均で 69.7% (76.9%)の記述子が除去された．

表 2. 論文誌名と文献数(ACM)

| |
|---|
| Trans. on Computer-Human Interaction (20), |
| Trans. on Computer Systems (60), |
| Trans. on Database Systems (165), |
| Trans. on Information Systems (105) |
| Trans. on Software Engineering and Methodology (50). |

表 3. 論文誌名と文献数(OMD)

| |
|---|
| American Heart Journal (50), |
| American Journal of Cardiology (50), |
| American Journal of Gastroenterology (50), |
| American Journal of Human Genetics (50), |
| American Journal of Obstetrics and Gynecology (50), |
| Anesthesiology (50), Blood (50), Cancer (50), |
| Chest (50), Circulation (50). |

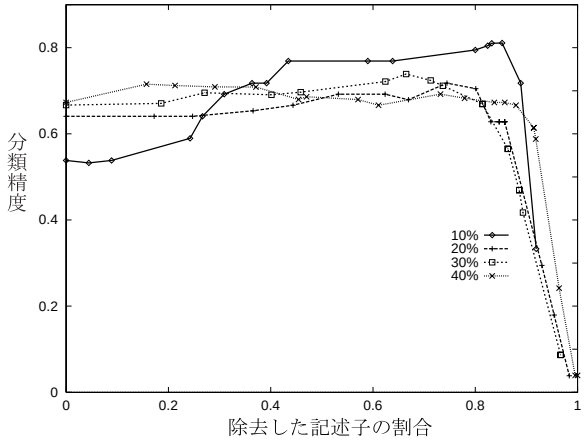


図 2. 分類精度の変化(ACM)

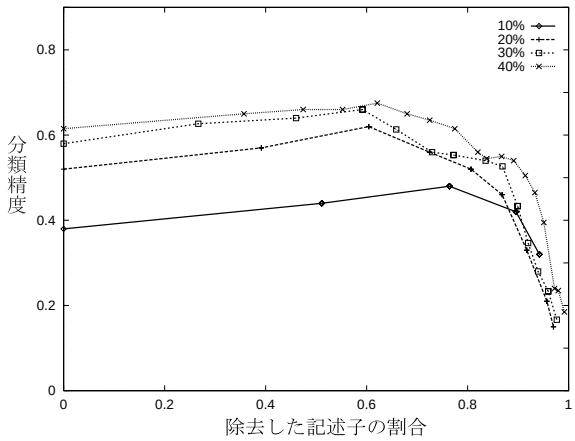


図 3. 分類精度の変化(OMD)

残った記述子をもとに，レベル 2 での文献クラスタの表現 $G(C_j,2)$から重みの低い順に記述子を除去したときの分類精度の変化を見た．分類対象と同じクラスタに属する他の文献の数を k として k-NN 法を利用した．結果を図 2 と 3 に示す．横軸は除去したレベル 2 での記述子より特定的な $d_i^c$ の記述子の数を初期の記述子の数で割った値である．縦軸は，分類精度である．コレクションとサイズの違いによらず，記述子の重み順に除去していくと，精度が徐々に

飽和した後低下するという曲線になった．結果，手続き 1 の終了時に ACM (OMD)では各 $G(C_j,2)$から平均で 89.5%（93.5%）の記述子が除去された．

次に，テスト用コレクションの文献を対象に手続き 2 での分類精度を調べた．結果を平均で表 4 と 5 に示す（提案手法と記す．括弧中の数字は，分類段階で参照した各クラスタ表現にある記述子の数の平均である）．k-NN 法での k は 10 とした．また，前処理としての文献の表現の簡略化を行わなかった場合（前処理なし），表現の簡略化を全く行わなかった場合（学習なし）の結果も示した．さらに，キーワードを概念階層中の記述子に置き換えずそのまま文献の表現に用い，類似度には重みを考慮したコサイン関数 [13] を利用して，k-NN 法を用いて分類を行った場合（k-NN）と， SVM を用いた場合の結果（SVM）も載せてある（括弧内の数字は，各カテゴリごとの文献のユニークなキーワードの数の平均である）．SVM の結果は，複数クラスタに対応したソフトウェア BSVM [5] を用いたときの，最良の値を載せてある．

ACM では，学習用コレクションの割合に関わらず，他の方法よりも同じか良い結果が得られた．学習段階で，不必要な記述子がうまく除去できたと言える．OMD については，割合が 30%，40%の場合に上と同様に優れていた．概念的表現ではキーワードが非常に多くの記述子に置き換えられたにも関わらず，不必要な記述子がうまく除かれたことがわかる． OMD では内容をうまく表す多くの専門的キーワードが文献に与えられたことから，学習事例が少ない場合に SVM がやや優れた結果となった．ただし，クラスタの内容をキーワードの羅列により理解することは難しい．

次に，簡略化されたクラスタの概念的表現で内容や他のクラスタとの関連が明らかになっているかを調べた．図 4 に全体の 40%を学習用コレクションとした場合に得られた ACM と OMD の各クラスタの表現から重みの大きい記述子を示す(OMD では 5 つのクラスタについて示した)．

図 4(a)から, TOCHI（Trans. on Computer-Human Interaction の省略．以下同様）と TOIS は“ software engineering”と“ information interfaces ”に関する文献を共通に含み，特に TOIS は，“ information storage and retrieval ”に関する文献も含むことが分かる．TOCS と TOSEM では，両者とも“ software engineering ”に関する文献を含みながら, 前者は，“ computer-communication networks ”などハードウェアに関するものを，後者は，“ programming languages ”などソフトウェアに関するものを多く含む．TODS のほとんどの文献が，“ database management ”に関連したものであった．

図 4(b) からは，AmHeartJ （American Heart Journal の省略，以下同様）， AmJCardiol, Circulation の 3 つは”heart diseases”に関連したものとわかる．Circulation では，他の 2 つが“ diagnostic techniques and procedures ”に関する文献を共通に多く含むのに対し，“ cardiovascular surgical procedures ” に関しての文献を多く含む．残り 2 つのクラスタは他と異なり,“ proteins ”に関する文献を共通に含む．また，Blood は，免疫，AmJHumGenet は遺伝子に関わる文献を多く含むことが分かる．以上のように，ここでの方法により，クラスタの内容や他との関連を容易に明らかにできることが言える．

表 4. 分類精度(ACM)

| | 学習用コレクションの割合 | | | |
|---|---|---|---|---|
| | 10% | 20% | 30% | 40% |
| 提案手法 | 0.69 (1.6) | 0.73 (2.6) | 0.74 (8.6) | 0.77 (9.6) |
| 前処理なし | 0.69 (9.4) | 0.73 (10.6) | 0.70 (9.4) | 0.74 (13.6) |
| 学習なし | 0.63 (46.6) | 0.67 (50.6) | 0.71 (53.8) | 0.72 (54.6) |
| k-NN | 0.64 (85.4) | 0.73 (135.0) | 0.74 (168.0) | 0.72(195.8) |
| SVM | 0.68 (85.4) | 0.71 (135.0) | 0.71 (168.0) | 0.75 (195.8) |

表 5. 分類精度(OMD)

| | 学習用コレクションの割合 | | | |
|---|---|---|---|---|
| | 10% | 20% | 30% | 40% |
| 提案手法 | 0.49 (6.5) | 0.57 (10.1) | 0.61 (47.7) | 0.62 (32.6) |
| 前処理なし | 0.43 (18.3) | 0.54 (30.2) | 0.57 (69.5) | 0.56 (84.6) |
| 学習なし | 0.48 (227.7) | 0.57 (327.0) | 0.60 (379.5) | 0.62 (438.8) |
| k-NN | 0.48 (110.0) | 0.56 (186.9) | 0.60 (259.1) | 0.62 (330.7) |
| SVM | 0.50 (110.0) | 0.59 (186.0) | 0.61 (259.1) | 0.62 (330.7) |

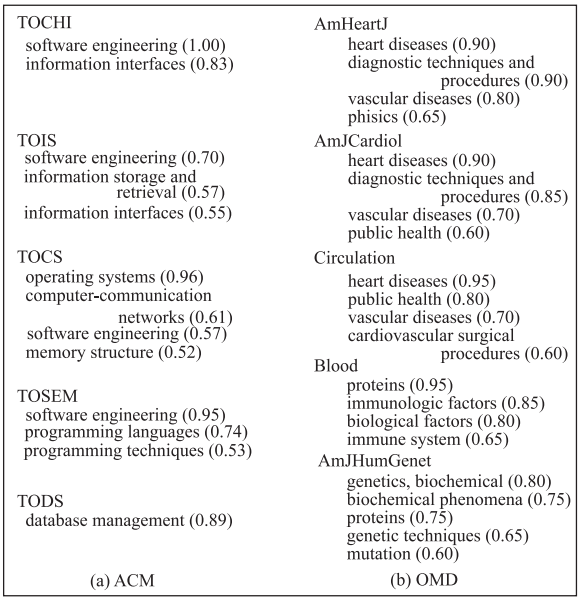


図 4. 簡略化されたクラスタの概念的表現

## 7. まとめ

分類精度を維持したまま，文献クラスタの内容を概念的に捉えられる簡潔な表現を得られる自動分類の方法を提案した．学習結果であるクラスタの表現からは他のクラスタとの関連も明らかになった．

クラスタの簡潔な表現を提示することは，新たな文献を分類しようとする場合に，適切なクラスタを確認するための指標となる．一方で，ある文献を分類したいクラスタに含めるには，内容的に何に関する記述が欠如しているかといった示唆を与えることにもなる．また，クラスタ間の関連が明らかになることから，関連の深いクラスタ同士を統合し，階層的なクラスタの管理と文献の分類を行えるようになると考えられる．

今後の課題として，$G(Cj,l)$ の簡略化を行う際，概念階層の特定のレベルに注目したが何れのレベルが適切かを決める手順が要る．定式化した方法の汎用性をより詳しくみるには，実験で使った以外のコレクションで有効性を確かめる必要がある．また，クラスタの概念的表現に人手による文献の分類結果をフィードバックさせるなど，精度の向上を目指すことも考えられる．

## 参考文献


[1] Association for Computing Machinery. ACM Computing Classification System [On-line]. 1997. Available: http://www.acm.org/class/1991

[2] Association for Computing Machinery (ACM). The ACM Digital Library [On-line]. 2002. Available: http://www.acm.org/dl

[3] S.T. Dumais and H. Chen. Hierarchical classification of web content. In Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR2000), pp. 256-263, 2000.

[4] S. T. Dumais, E Cutrell, and H. Chen. Optimizing search by showing results in context. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'01), pp. 273-283, 2001.

[5] C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. IEEE Trans. on Neural Networks, vol. 13, pp. 415-425 2002.

[6] Excite Japan Co. Ltd. Excite [On-line]. http://www.excite.co.jp

[7] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In Proceedings of the 10th European Conference on Machine Learning (ECML-98), number 1398, pp. 137-142, 1998.

[8] W. Lam and C. Y. Ho. Using a generalized instance set for automatic text categorization. In Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), pp. 81-89, 1998.

[9] 中島誠, 伊藤哲郎. 質問との概念的関連性を捉えるための文献内容表現の扱い, 電子情報通信学会論文誌, vol. J85-D-I, no.5, pp. 436-444, 2002.

[10] National Library of Medicine. Medical Subject Headings [On-line]. 1998. Available: http://www.nlm.nih.gov/mesh/meshhome.html

[11] Oregon Health Sciences University, ftp://medir.ohsu.edu/pub/ohsumed/

[12] M.F. Porter, An algorithm for suffix stripping. Program, vol. 14, no. 3, pp. 130-137, 1980.

[13] G. Salton and M. McGill. Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983.

[14] C. J. van Rijsbergen. Information retrieval. Butterworths, London, 1979.

[15] Yahoo! Inc. Yahoo! [On-line]. http://www.yahoo.com/

[16] Y. Yang. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), pp. 13-22, 1994.

[17] Y. Yang and X. Liu. A re-examination of text categorization methods. In Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), pp. 42-49, 1999.